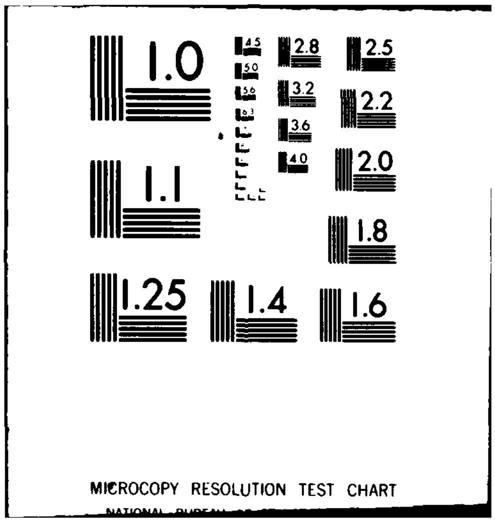
AD-A089 643 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
THE DUALITY OF DIAGNOSTIC CHECKING AND ROBUSTIFICATION IN MODEL--ETC(U)
JUN 80 S P BAILEY, G E BOX DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-2086 NL

END
DATE
FILMED
10-80
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART

(2)

AD A089643

(9) MRC Technical Summary Report #2086

(6) THE DUALITY OF DIAGNOSTIC CHECKING AND ROBUSTIFICATION IN MODEL BUILDING: SOME CONSIDERATIONS AND EXAMPLES.

(10) Steven P. Bailey and
George E. P. Box

(14) MRC-TSR-2086

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

(11) (12) 54

June 1980

(Received April 14, 1980)

(15) DAAG29-75-C-0024, DAAG29-80-C-0041

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

THE DUALITY OF DIAGNOSTIC CHECKING AND ROBUSTIFICATION IN MODEL BUILDING: SOME CONSIDERATIONS AND EXAMPLES

Steven P. Bailey and George E. P. Box

Technical Summary Report #2086
June 1980

ABSTRACT

Consideration is given to the means by which appropriate diagnostic checking functions of the data can be developed to guard against feared model discrepancies. A formal basis for the selection of a function is given for situations where the feared inadequacy can be characterized by a discrepancy parameter β (beta) which takes a (possible inappropriate) value of β_0 (beta sub-zero) in the model. The relationship of this checking function with the posterior distribution obtained from an elaborated ("robustified") model which allows for the discrepancy parameter to be estimated is discussed. The nature of the diagnostic check is briefly described for problems relating to transformation of the dependent variable and to serial correlation; while a more thorough investigation of the checking function is given for problems relating to outlying observations and to transformation of predictor variables. Several examples are given to illustrate these ideas.

AMS(MOS) Subject Classification - 62G35

Key Words: Diagnostic checks, robustification, discrepancy parameters, outliers, transformations, serial correlation.

Work Unit No. 4 - Statistics and Probability

Sponsored by the Unites States Army under Contract Nos. DAAG29-75-C-0024 and DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

Statistical methods are useful as tools in scientific model-building investigations. In particular, at each stage of an investigation it is of interest to not only estimate the parameters of the model being postulated at that time, but to also check the fitted model in its relation to the data with the intent to reveal its inadequacies, if any.

In this paper, a diagnostic checking function is developed for the latter purpose above. This diagnostic check is useful in situations where a feared model inadequacy can be characterized by a so-called discrepancy parameter β which may have a true value different from the value β_0 assumed by the current model. The relationship of using this checking function to using a broader model which allows β to be estimated (rather than assuming that $\beta = \beta_0$) is explored.

Examples are discussed in which β measures the need to allow for outliers, the need for data transformation, and the need to allow for serial correlation of errors.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

THE DUALITY OF DIAGNOSTIC CHECKING AND ROBUSTIFICATION IN MODEL BUILDING: SOME CONSIDERATIONS AND EXAMPLES

Steven P. Bailey and George E. P. Box

1. Introduction.

"The advancement of statistical methods to the present state has depended critically upon the interaction of mathematics with real data, and we cannot but benefit if more attention is given to the characteristics of the real world than has been done in recent decades." This view, expressed by Stigler (1977), is characteristic of the recent renewal of interest in finding out what the real world is really like. (See also Box, 1979d.)

In this spirit, Chen and Box (1979a,b) recently conducted a study of real data. In their investigation of nine data sets, they used the contaminated exponential power distribution in modeling the experimental errors, and the results suggest, in their words, that "heavy tailed distributions are sometimes caused by inhomogeneity in mean and variance such as might be encountered early in an experiment because of start-up difficulties. After such inhomogeneity has been allowed for, the observations seem to be adequately represented by a contaminated normal distribution. Thus, for a carefully planned experiment, a contaminated normal distribution is likely to be appropriate."

Thus a particular combination of robustification and diagnostic checking techniques are being recommended as

Sponsored by the United States Army under Contract Nos. DAAG29-75-C-0024 and DAAG29-80-C-0041.

appropriate for modeling data from a well-planned investigation. Specifically, the investigator who wishes to implement the above suggestions can do so by making provisions in the model for the possibility of discrepant observations (and thereby robustifying against outliers) while tentatively excluding from the model additional provisions for the possibility of non-normal kurtosis (and thereby being content to later perform a diagnostic check to assess whether or not this exclusion is warranted).

Of course, the experienced investigator will be aware of the fact that there are numerous other potential problems besides the two mentioned above which need to be considered when deciding upon what to include in a model versus what to tentatively omit from a model. Of particular importance are questions concerning the possibility of serial correlation and the need to transform the response and/or predictor variables.

Since it is impossible to attempt to provide for all possible contingencies when choosing a model, it is important for the investigator to have available appropriate diagnostic checking procedures which are capable of detecting when deficiencies, such as the ones mentioned above, exist in the model. A general method for developing such procedures will now be considered.

2. A formal basis for the selection of a diagnostic checking function.

As discussed by Box (1979b), criticism of a tentative model M in the light of the observed data $\underset{\sim}{y}_d$ is often done informally (for example, by examining residuals), and yet has an underlying formal justification in that any peculiar aspect of $\underset{\sim}{y}_d$ which may cause the investigator to doubt the adequacy of M will correspond to an appropriate summary measure $g(\underset{\sim}{y}_d)$ of that aspect which is judged unusual when assessed against its predictive reference distribution $p(g(\underset{\sim}{y})|M)$.

A formal basis for the selection of a checking function $g(\underset{\sim}{y})$ has recently been suggested by Box (1979c) in the context of considering any particular model departure that can be represented in terms of a so-called discrepancy parameter β. Letting β_0 denote the assumed value of β for the model M, this proposed checking function is given by

$$g_\beta(\underset{\sim}{y}) = \frac{\partial}{\partial\beta} \ln p(\underset{\sim}{y}|\beta)\Big|_{\beta = \beta_0} , \qquad (1)$$

where $p(\underset{\sim}{y}|\beta)$ denotes the conditional predictive distribution for a given choice of β (and hence $p(\underset{\sim}{y}|\beta_0)$ is the predictive distribution appropriate for the model M).

One justification for the appropriateness of (1) as a diagnostic checking function can be developed by noting the relationship between the conditional predictive

distribution $p(\underset{\sim}{y}|\beta)$ and the pseudo-likelihood function $p_\ell(\beta|\underset{\sim}{y})$. As used by Box (1979a; see also Bailey and Box, 19 the pseudo-likelihood arises when, instead of setting β at a specified value β_0 for the model, the discrepancy parameter is assessed a prior distribution $p(\beta)$ and estimated along with the other model parameters. Defined by

$$p_\ell(\beta|\underset{\sim}{y}) = \frac{p(\beta|\underset{\sim}{y})}{p(\beta)} , \tag{2}$$

the pseudo-likelihood differs for differing choices of $p(\beta)$ only by a multiplicative constant which depends on $\underset{\sim}{y}$, and hence it represents information about β contained in the data. It is convenient to consider for definiteness a standardized form of (2) ; namely, the posterior distribution $p_u(\beta|\underset{\sim}{y})$ which results from using a uniform prior $p_u(\beta)$. For this choice it then follows from

$$p(\beta|\underset{\sim}{y}) \propto p(\underset{\sim}{y}|\beta)p(\beta) \tag{3}$$

that

$$p_u(\beta|\underset{\sim}{y}) \propto p(\underset{\sim}{y}|\beta) . \tag{4}$$

Hence (1) can be rewritten

$$g_\beta(\underset{\sim}{y}) = \frac{\partial}{\partial\beta} \ln p_u(\beta|\underset{\sim}{y})\Big|_{\beta = \beta_0} \qquad (5)$$

so that this predictive check is seen to be measuring an intuitively interesting quantity; namely, the instantaneous rate of change of the logarithm of the pseudo-likelihood (or "log pseudo-likelihood", for short) at the "ideal" value β_0 .

The interpretation of the predictive check (1) will depend upon the nature of the discrepancy parameter β under study. Two different situations can occur.

First, consider the case where it makes sense to talk about possible β values over a wide range, with the "ideal" value β_0 somewhere in the middle of this range. Thus, if the actual data observed, $\underset{\sim}{y}_d$, give rise to a value $g_\beta(\underset{\sim}{y}_d)$ which is small in absolute value, then the appropriateness of an assumption $\beta = \beta_0$ is not seriously questioned. However, an unusually large value of $|g_\beta(\underset{\sim}{y}_d)|$, as assessed by the predictive distribution $p(g_\beta(\underset{\sim}{y}_d)|\beta_0)$, will cast doubt on the validity of the $\beta = \beta_0$ model assumption. Specifically, a large positive value of $g_\beta(\underset{\sim}{y}_d)$ will suggest that values of β that are larger than β_0 may be more reasonable to consider, whereas a large negative value of $g_\beta(\underset{\sim}{y}_d)$ will suggest that smaller values of β than the ideal value β_0 may be more appropriate. Hence, the use of this predictive check corresponds, in a sampling theory

framework, to a hypothesis test of $\beta = \beta_0$ against a two-sided alternative. Several examples will be given later in this section.

Now, consider the case* where it only makes sense to talk about β values for which $\beta \geq \beta_0$. Here, the predictive check (1) can be viewed as a sampling theory hypothesis test of $\beta = \beta_0$ against the one-sided alternative corresponding to larger values of β. (An example relevant to this case will be presented in Section 3.)
smaller values of $g_\beta(\underset{\sim}{y})$ will lend more support to a model assumption of $\beta = \beta_0$ than will larger values of $g_\beta(\underset{\sim}{y})$. Of course a convenient reference value to compare $g_\beta(\underset{\sim}{y})$ to in this case is zero since, from the earlier argument, a value $g_\beta(\underset{\sim}{y}_d) > 0$ will indicate that the log pseudo-likelihood will initially increase as β is allowed to increase from β_0, while a value $g_\beta(\underset{\sim}{y}_d) < 0$ will indicate an initial decrease in $p_u(\beta|\underset{\sim}{y})$ when β is increased from β_0.

<u>Some applications of the proposed checking function</u>.

Box (1979c) illustrates the use of the checking function (1) by applying it in two particular situations which we now briefly discuss.

---

*An analogous argument can, of course, be given for the case where $\beta \leq \beta_0$ by redefining the discrepancy parameter to be the negative of what it presently is and then using the above development.

Suppose that the investigator tentatively fits, to the n-dimensional vector $\underset{\sim}{y} = (y_1, \ldots, y_n)'$ of untransformed observations, the normal linear model $N_n(X\underset{\sim}{\theta}, \sigma^2 I)$, where $\underset{\sim}{\theta}$ is p-dimensional. However, suppose further that he wishes to perform a diagnostic check which would indicate whether or not it would be more appropriate to fit the same model to some suitable transformation of the observations. In particular, if the family of transformations

$$y_u^{(\lambda)} = \begin{cases} \dfrac{y_u^{\lambda} - 1}{\lambda} & , \quad \lambda \neq 0 \\ \ell n \; y_u & , \quad \lambda \neq 0 \end{cases} \qquad u = 1, \ldots, n \tag{6}$$

discussed by Box and Cox (1964) is considered, then a diagnostic checking function $g_\lambda(\underset{\sim}{y})$ of the form (1) can be used, where $\lambda_0 = 1$ is the value of λ for the tentative model.

For this situation, Box (1979c) derives the conditional predictive distribution

$$\begin{aligned} p(\underset{\sim}{y}|\lambda) &\propto \dot{y}^{\nu(\lambda-1)}[\underset{\sim}{y}^{(\lambda)}{}' R \; \underset{\sim}{y}^{(\lambda)}]^{-\frac{1}{2}\nu} \\ R &= I - X(X'X)^{-1}X' \quad , \quad \nu = n - p \quad , \end{aligned} \tag{7}$$

where $\dot{y}$ is the geometric mean of the untransformed

observations. Upon taking the logarithm of (7), differentiating with respect to λ, and evaluating this derivative at $\lambda_0 = 1$, he finds the resulting diagnostic checking function to be

$$g_\lambda(\underset{\sim}{y}) = \frac{\underset{\sim}{Y}'R\,\underset{\sim}{y}}{s^2} , \tag{8}$$

where

$$\left.\begin{aligned} \nu s^2 &= \underset{\sim}{y}'R\,\underset{\sim}{y} , \\ Y_u &= y_u[1 - \ln(\frac{y_u}{\dot{y}})] , \quad u = 1,\ldots,n . \end{aligned}\right\} \tag{9}$$

Thus, the predictive check (8) seeks a correlation between the residuals from fitting $\underset{\sim}{y}$ to $N_n(X\underset{\sim}{\theta},\sigma^2 I)$ and the residuals from fitting $\underset{\sim}{Y}$ to $N_n(X\underset{\sim}{\theta},\sigma^2 I)$, since the two residual vectors are given by $R\underset{\sim}{y}$ and $R\underset{\sim}{Y}$, respectively. If a strong correlation is found, then this indicates the need for transformation. Now, by recalling the earlier discussion pertaining to the general predictive check $g_\beta(\underset{\sim}{y})$ in its form (5), it is seen that a strong positive correlation between these two sets of residuals will indicate that values of $\lambda > \lambda_0 = 1$ are worthy of consideration; whereas a strong negative correlation between the two residual sets is seen to indicate that values of $\lambda < \lambda_0 = 1$ may be more appropriate. An example

illustrating the application of this diagnostic check will be given in Section 4.

Note that this predictive check can be conducted in informal manner by simply plotting the two sets of residuals against each other. Note also the relationship of this check to the somewhat similar checks proposed by Tukey (1949) and by Andrews (1971). Specifically, by defining $Z_u^{(T)} = \hat{y}_u^2$ and $Z_u^{(A)} = \hat{y}_u \ln \hat{y}_u$, where the $\hat{y}_u$'s are the predicted values from a fit of $\underset{\sim}{y}$ to $N_n(X\theta,\sigma^2 I)$, it is then seen that Tukey's transformable nonadditivity check seeks a correlation between the residuals from a fit of the y_u 's to the model and the residuals *from a fit of the* $Z_u^{(T)}$'s to the same model. Similarly, Andrews' check is concerned with the correlation between the residuals from fitting the y_u 's and those from fitting the $Z_u^{(A)}$'s .

We now turn to a second illustration of the applicability of the general predictive check (1). Box (1979c) considers the situation in which the investigator tentatively assumes the independence of the y_u 's but wishes to perform a diagnostic check relative to the possible presence of first-order autoregressive behavior in the error structure. (See, for example, Pallesen, 1978.)

We shall not go through the details of deriving this result here. It turns out that the diagnostic checking function (1). is a multiple of the lag one sample autocorrelation coefficient,

upon which the familiar test of Durbin and Watson (1950, 1951) is based. As was the case with the predictive check corresponding to the need for transformation, this check for serial correlation can, if preferred, be performed in an informal manner through residual plotting (in this case, by plotting the u^{th} residual versus the $(u-1)^{st}$ residual for $u = 2,\ldots,n$).

3. A diagnostic check for discrepant observations.

We now illustrate the nature of the checking function (1) for the situation described by Box and Tiao (1968), where each observation has a probability α of being discrepant in some specific manner. The predictive distribution conditional on a specific choice of α is expressed in terms of the $n+1$ predictive distributions conditional on specific choices of the number of outliers r as

$$p(\underset{\sim}{y}|\alpha) = \sum_{r=0}^{n} \binom{n}{r}\alpha^r(1-\alpha)^{n-r}p(\underset{\sim}{y}|r)$$

(Bailey and Box, 1980), so that

$$\frac{\partial}{\partial\alpha}\ln p(\underset{\sim}{y}|\alpha) = \frac{\sum_{r=0}^{n}\left[\frac{\partial}{\partial\alpha}\alpha^r(1-\alpha)^{n-r}\right]\binom{n}{r}p(\underset{\sim}{y}|r)}{\sum_{r=0}^{n}\binom{n}{r}\alpha^r(1-\alpha)^{n-r}p(\underset{\sim}{y}|r)}$$

$$= \frac{-n(1-\alpha)^{n-1}p(\underset{\sim}{y}|r=0)+\{\sum_{r=1}^{n-1}[(n-j\alpha)\alpha^{j-1}(1-\alpha)^{n-j-1}]\binom{n}{j}p(\underset{\sim}{y}|r=j)\}+n\alpha^{n-1}p(\underset{\sim}{y}|r=n)}{\sum_{r=0}^{n}\binom{n}{r}\alpha^r(1-\alpha)^{n-r}p(\underset{\sim}{y}|r)} \quad (10)$$

Thus, the checking function $g_\alpha(\underset{\sim}{y})$, corresponding to departures from the ideal situation where $\alpha_0 = 0$, becomes

$$g_\alpha(\underset{\sim}{y}) = \frac{-n\ p(\underset{\sim}{y}|r=0)+n\ p(\underset{\sim}{y}|r=1)}{p(\underset{\sim}{y}|r=0)} = n\left[\frac{p(\underset{\sim}{y}|r=1)}{p(\underset{\sim}{y}|r=0)} - 1\right] = n(h_1 - 1) \ , \tag{11}$$

with $h_j = \dfrac{p(\underset{\sim}{y}|r=j)}{p(\underset{\sim}{y}|r=0)}$.

As a first application of this checking function, consider the analysis of Darwin's data presented in Bailey and Box (1980), which uses a "mistaken sign" model for the bad observations. Using the value $h_1 = 2.10$ from Table 5 of that paper results in $g_\alpha(\underset{\sim}{y}) = 15(2.10 - 1) = 16.5$. Recalling again the interpretation of this diagnostic check that is suggested by (5), this value thus indicates that $\ln p_u(\alpha|\underset{\sim}{y})$ will have an initial slope of 16.5 at $\alpha = \alpha_0 = 0$ (inspection of $p_u(\alpha|\underset{\sim}{y})$ as given in Figure 10a of that paper bears this out) and hence indicates the relative implausibility of an assumption that $\alpha = 0$ (that is, an assumption that, with probability one, none of the 15 observations has an incorrect sign).

Now, $r = 0$ and $r = 1$ are the only cases which enter into the calculation of $g_\alpha(y)$. For convenience, we introduce the following notation. Let the subscript "i" denote those quantities which pertain to an assumption that

the i^{th} observation is bad and all other observations are good. Then, for example, $\underset{\sim}{z}_i$ would be a vector of zeroes, except for the i^{th} element, which would equal one and would thus single out the i^{th} observation as being discrepant. Also, let $i=0$ correspond to the assumption that no observations are discrepant.

Returning, then, to the consideration of $g_\alpha(\underset{\sim}{y})$, note that

$$p(\underset{\sim}{y}|r=1) = \frac{1}{n}\sum_{i=1}^{n} p(\underset{\sim}{y}|\underset{\sim}{z}_i), \tag{12}$$

so that the predictive check (11) can be rewritten

$$g_\alpha(\underset{\sim}{y}) = \sum_{i=1}^{n} g_{i,\alpha}(\underset{\sim}{y}) , \tag{13}$$

where
$$g_{i,\alpha}(\underset{\sim}{y}) = \frac{p(\underset{\sim}{y}|\underset{\sim}{z}_i)}{p(\underset{\sim}{y}|\underset{\sim}{z}_0)} - 1 . \tag{14}$$

Thus (13) expresses the checking function $g_\alpha(\underset{\sim}{y})$ as a summation of n individual checking functions, the i^{th} of which compares the predictive ratio of the hypotheses "only y_i discrepant" and "no observations discrepant" to unity.

It is of interest to consider the nature of this diagnostic check for the situation described in Bailey and Box (1980) where $p(\underset{\sim}{y}|\underset{\sim}{\theta},\sigma^2,\underset{\sim}{z}_i)$ is $N_n(X_i\underset{\sim}{\theta},\sigma^2\Sigma_i)$ for $i = 0,1,\ldots,n$. Using their results, it follows that (14) reduces to

$$g_{i,\alpha}(\underset{\sim}{y}) = \left[\frac{|\Sigma_i||X_i'\Sigma_i^{-1}X_i|}{|\Sigma_0||X_0'\Sigma_0^{-1}X_0|}\left(\frac{s_i^2}{s_0^2}\right)^{\nu}\right]^{-\frac{1}{2}} - 1 , \tag{15}$$

where

$$\left.\begin{aligned} p &= \text{length of parameter vector } \underset{\sim}{\theta}, \\ \nu &= n-p, \\ \hat{\underset{\sim}{\theta}}_i &= (X_i'\Sigma_i^{-1}X_i)^{-1}X_i'\Sigma_i^{-1}\underset{\sim}{y}, \\ \nu s_i^2 &= (\underset{\sim}{y}-X_i\hat{\underset{\sim}{\theta}}_i)'\Sigma_i^{-1}(\underset{\sim}{y}-X_i\hat{\theta}_i). \end{aligned}\right\} \quad (16)$$

In particular, if it is assumed that a bad observation is one having a standard deviation which is k times as large as the common standard deviation of all good observations (see, for example, Box and Tiao, 1968), then (15) becomes

$$g_{i,\alpha}(y) = \frac{1}{k}\left[\frac{|X_0'X_0-\phi\underset{\sim}{x}_i\underset{\sim}{x}_i'|}{|X_0'X_0|}\right]^{-\frac{1}{2}}\left[1+\frac{c_i}{\nu-1}r_i^2\right]^{\frac{\nu}{2}}-1 \quad (17)$$

where

$$\left.\begin{aligned} X_i &= X_0 \text{ for all } i = 1, \ldots, n \\ \underset{\sim}{x}_i' &= i^{th} \text{ row of } X_0, \\ \phi &= 1-\frac{1}{k^2}, \\ c_i &= \left(\frac{\nu-1}{\nu}\right)\phi(1-\phi\underset{\sim}{x}_i'(X_0'X_0)^{-1}\underset{\sim}{x}_i), \\ r_i &= \frac{y_i-\underset{\sim}{x}_i'\hat{\underset{\sim}{\theta}}_i}{s_i}. \end{aligned}\right\} \quad (18)$$

For the special case, with p=1, where θ is a location parameter, and thus $X_0 = \underset{\sim}{1}$, it follows that (17) reduces to

$$g_{i,\alpha}(\underset{\sim}{y}) = \frac{1}{k}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}} R_i - 1, \tag{19}$$

where

$$R_i = \left(1 + \frac{\phi}{n-1}\frac{n-\phi}{n} r_i^{\,2}\right)^{\frac{n-1}{2}} . \tag{20}$$

Hence, from (18), the predictive checking function is

$$g_{\alpha}(\underset{\sim}{y}) = \frac{1}{k}\left(\frac{n}{n-\phi}\right)^{\frac{1}{2}} R - n, \tag{21}$$

where

$$R = \sum_{i=1}^{n} R_i . \tag{22}$$

Note that the R_i's and R quantities considered by Chen and Box (1979a) in their investigation of the weighting structure of the posterior mean of a location parameter for a contaminated normal population. Specifically, each R_i is proportional to the inverse of the t-ordinate corresponding to the standardized residual r_i of the "bad" observation, with r_i defined in (18); so that if r_i is large, due to the i^{th} observation truly being discrepant, then R_i and thus R will also be large, yielding an unusually large value of the predictive checking function $g_{\alpha}(\underset{\sim}{y})$, as given by (21).

We now turn to a specific example which illustrates the practical use of the diagnostic checks discussed in these last two sections.

4. An example: John's 2^5 data.

John (1978) gives data from a 2^5 factorial experiment designed to study the effects of five factors on the strength of a type of metal coating material, with abrasion loss as the measured response. The experiment was run in two blocks, confounding the highest order interaction with the operator effect, as two different workers applied the coating.

The experimental design is set out in Table 1a, where x_1-x_5 and x_6 denote, respectively, the five factors under study and the blocking factor, and where for each x_i the coding is such that the two factor levels used in the experiment correspond to $x_i=-1$ and $x_i=1$. The data generated from this design are given in Table 1b, along with the corresponding predicted values and residuals from fitting the data to the model

$$y_u = \theta_0 + \sum_{i=1}^{5} \theta_i x_{iu} + \sum_{i=1}^{4} \sum_{j=i+1}^{5} \theta_{ij} x_{iu} x_{ju} + \theta_6 x_{6u} + \varepsilon_u; \; u=1, \ldots, 32, \qquad (23)$$

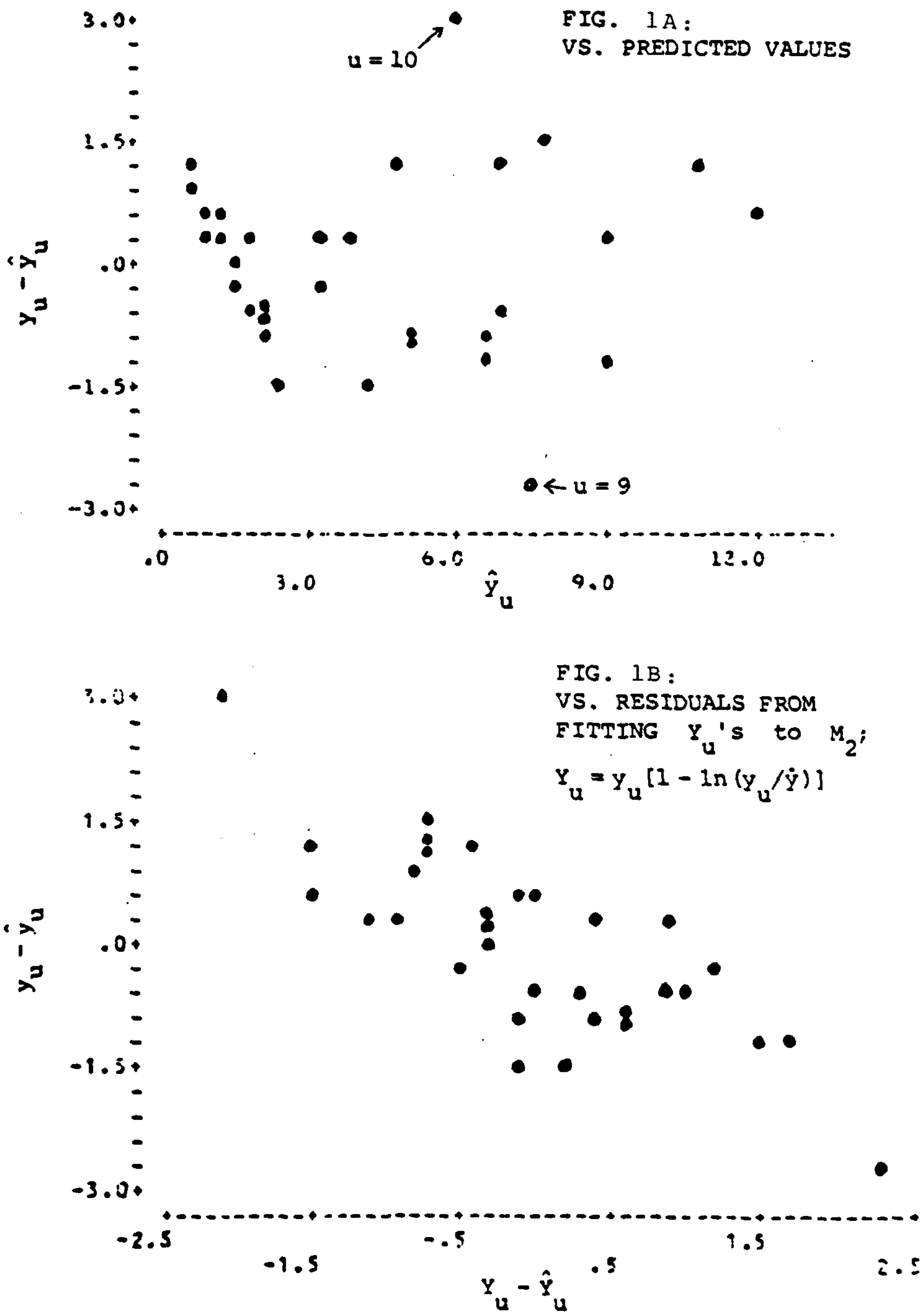
where x_{iu} denotes the value of x_i corresponding to the u^{th} observation.

These residuals and predicted values are plotted against each other in Figure 1a, and inspection of this plot shows two possibly deviant observations (y_{10} and y_9). John performs a significance test for two outliers, based upon the work of John and Draper (1978), and finds "no evidence of any outliers being present". However, he tempers this conclusion by noting that a similar significance test for just one outlier does give indication that y_{10} is an outlier.

TABLE 1. RESULTS FROM CONFOUNDED 2^5 EXPERIMENT DISCUSSED BY JOHN.

u	(a) The design x_{1u}	x_{2u}	x_{3u}	x_{4u}	x_{5u}	x_{6u}	(b) Raw data y_u: y_u	$\hat{y}_u$	$y_u - \hat{y}_u$	Fitting M_2 to (c) $Y_u = y_u[1 - \ln(y_u/\dot{y})]$: Y_u	$\hat{Y}_u$	$Y_u - \hat{Y}_u$	(d) $y_u^{(0)} = \ln y_u$: $y_u^{(0)}$	$\hat{y}_u^{(0)}$	$y_u^{(0)} - \hat{y}_u^{(0)}$
1	-1	-1	-1	-1	-1	-1	1.4	1.11	.29	2.67	3.49	-.92	.34	.49	-.15
2	1	-1	-1	-1	-1	1	1.2	1.89	-.69	2.39	2.13	.26	.18	.33	-.15
3	-1	1	-1	-1	-1	1	3.6	3.26	.34	3.20	2.26	.94	1.28	.93	.35
4	1	1	-1	-1	-1	-1	1.2	1.51	-.31	2.39	2.88	-.50	.18	.43	-.25
5	-1	-1	1	-1	-1	1	1.5	.54	.96	2.65	3.41	-.77	.41	.24	.16
6	1	-1	1	-1	-1	-1	1.4	1.99	-.59	2.57	1.61	.96	.34	.35	-.01
7	-1	1	1	-1	-1	-1	1.5	2.14	-.64	2.65	2.63	.01	.41	.61	-.21
8	1	1	1	-1	-1	1	1.6	.96	.64	2.72	2.71	.01	.47	.22	.25
9	-1	-1	-1	1	-1	1	5.0	7.59	-2.59	2.80	.46	2.35	1.61	1.92	-.31
10	1	-1	-1	1	-1	-1	9.0	5.89	3.11	-.24	1.89	-2.14	2.20	1.65	.55
11	-1	1	-1	1	-1	-1	12.0	10.94	1.06	-3.77	-2.24	-1.54	2.48	2.49	-.01
12	1	1	-1	1	-1	1	5.4	6.61	-1.21	2.61	1.07	1.54	1.69	1.73	-.04
13	-1	-1	1	1	-1	-1	4.2	5.11	-.91	3.09	2.48	.60	1.44	1.58	-.15
14	1	-1	1	1	-1	1	4.4	3.99	.41	3.03	3.37	-.34	1.48	1.42	.06
15	-1	1	1	1	-1	1	9.3	7.81	1.49	-.55	.13	-.68	2.23	1.93	.30
16	1	1	1	1	-1	-1	2.8	4.16	-1.36	3.19	2.99	.20	1.03	1.43	-.40
17	-1	-1	-1	-1	1	1	1.7	1.64	.06	2.79	3.04	-.25	.53	.58	-.05
18	1	-1	-1	-1	1	-1	2.0	1.79	.21	2.95	2.59	.37	.69	.48	.21
19	-1	1	-1	-1	1	-1	3.1	3.34	-.24	3.22	2.04	1.18	1.13	.94	.20
20	1	1	-1	-1	1	1	1.2	.86	.34	2.39	3.46	-1.08	.18	.35	-.17
21	-1	-1	1	-1	1	-1	1.9	1.31	.59	2.90	2.96	-.06	.64	.41	.23
22	1	-1	1	-1	1	1	1.2	2.04	-.84	2.39	1.97	.42	.18	.43	-.25
23	-1	1	1	-1	1	1	1.0	2.36	-1.36	2.17	2.31	-.14	.00	.54	-.54
24	1	1	1	-1	1	-1	1.8	.56	1.24	2.85	3.29	-.44	.59	.22	.37
25	-1	-1	-1	1	1	-1	9.5	9.11	.39	-.77	-.51	-.26	2.25	2.24	.01
26	1	-1	-1	1	1	1	5.9	6.69	-.79	2.33	1.73	.60	1.77	1.88	-.11
27	-1	1	-1	1	1	1	12.6	11.91	.69	-4.58	-3.08	-1.50	2.53	2.58	-.05
28	1	1	-1	1	1	-1	6.3	6.96	-.66	2.08	1.14	.94	1.84	1.88	-.04
29	-1	-1	1	1	1	1	8.0	6.79	1.21	.73	1.41	-.69	2.08	1.83	.25
30	1	-1	1	1	1	-1	4.2	5.04	-.84	3.09	3.21	-.12	1.44	1.73	-.30
31	-1	1	1	1	1	-1	7.7	9.04	-1.34	.99	-.72	1.71	2.04	2.09	-.05
32	1	1	1	1	1	1	6.0	4.66	1.34	2.27	2.96	-.68	1.79	1.51	.29

FIGURE 1. RESIDUAL PLOTS FOR JOHN'S 2^5 DATA (UNTRANSFORMED) USING M_2.



When applying the predictive check (11) to this example, using k=5, it is found that h_1=1.90 (and, incidentally, h_2=2.03), so that $g_\alpha(\underset{\sim}{y})$ = 32[1.90-1] = 28.8. Note, in particular, that the value of this checking function is positive, which indicates that the log pseudo-likelihood for α will initially increase as α is allowed to increase from α_0=0.

Thus this diagnostic check is giving some indication that the basic model should be elaborated to allow for the possibility of outlying observations. This can be done using the methodology developed by Bailey and Box (1980), which entails specifying a prior distribution for frequency α in which discrepant observations occur. However for present purposes, it will be sufficient to consider a sensitivity analysis approach (Box and Tiao, 1968), whereby (i) an analysis of the data is performed for each of a number of different fixed values for α and (ii) these analyses are compared to assess how inferences are affected by the choice of α.

Thus, consider the model (23), where now the errors ε_u are assumed to come from a contaminated normal distribution for which (i) the nature of the contamination is an inflation of the variance by a factor of k^2=25 and (ii) the frequency of the contamination is given, in percentages, by 100α%. The posterior means and posterior standard deviations of the parameters in (23), based upon the assumption just given, are exhibited in Table 2 for various choices of α. Of course,

TABLE 2. POSTERIOR MEANS AND STD. DEVIATIONS (IN PARENTHESES) FOR θ's OF M_2 FOR JOHN'S 2^5 DATA (UNTRANSFORMED) FOR VARIOUS VALUES OF α WITH $k = 5$.

α	0	.001	.005	.01	.025	.05	.10
θ_0	4.36 (.31)	4.36 (.31)	4.33 (.30)	4.31 (.30)	4.28 (.29)	4.26 (.28)	4.25 (.27)
θ_1	-.89 (.31)	-.90 (.31)	-.92 (.30)	-.95 (.30)	-1.00 (.28)	-1.03 (.27)	-1.05 (.26)
θ_2	.46 (.31)	.46 (.31)	.49 (.30)	.51 (.30)	.54 (.29)	.56 (.28)	.57 (.27)
θ_3	-.71 (.31)	-.70 (.31)	-.68 (.30)	-.66 (.30)	-.62 (.29)	-.60 (.28)	-.59 (.27)
θ_4	2.66 (.31)	2.65 (.31)	2.63 (.30)	2.61 (.30)	2.57 (.29)	2.55 (.28)	2.54 (.27)
θ_5	.27 (.31)	.28 (.31)	.30 (.30)	.32 (.30)	.35 (.29)	.37 (.28)	.39 (.27)
θ_{12}	-.64 (.31)	-.64 (.31)	-.61 (.30)	-.58 (.30)	-.53 (.28)	-.50 (.27)	-.48 (.26)
θ_{13}	.16 (.31)	.16 (.31)	.19 (.30)	.22 (.30)	.26 (.28)	.29 (.27)	.31 (.26)
θ_{14}	-.63 (.31)	-.64 (.31)	-.67 (.30)	-.69 (.30)	-.74 (.28)	-.77 (.27)	-.80 (.25)
θ_{15}	-.17 (.31)	-.16 (.31)	-.13 (.30)	-.11 (.30)	-.06 (.28)	-.03 (.27)	-.01 (.26)
θ_{23}	-.15 (.31)	-.16 (.31)	-.18 (.30)	-.20 (.30)	-.23 (.29)	-.25 (.28)	-.27 (.27)
θ_{24}	.29 (.31)	.29 (.31)	.32 (.30)	.34 (.30)	.37 (.29)	.39 (.28)	.41 (.23)
θ_{25}	-.13 (.31)	-.13 (.31)	-.15 (.30)	-.17 (.30)	-.21 (.29)	-.23 (.28)	-.25 (.23)
θ_{34}	-.49 (.31)	-.48 (.31)	-.46 (.30)	-.44 (.30)	-.41 (.29)	-.39 (.28)	-.37 (.27)
θ_{35}	.05 (.31)	.04 (.31)	.02 (.30)	.00 (.30)	-.03 (.29)	-.05 (.28)	-.06 (.27)
θ_{45}	.24 (.31)	.24 (.31)	.27 (.30)	.29 (.30)	.32 (.29)	.34 (.28)	.36 (.27)
θ_6	-.01 (.31)	-.00 (.31)	.02 (.30)	.05 (.30)	.10 (.28)	.13 (.27)	.15 (.26)

the $\alpha=0$ column corresponds to an assumption that there are no discrepant observations. The other choices of α considered were $\alpha = .001, .005, .01, .025, .05$ and $.10$.*

A study of how the posterior mean change as α increases shows that there is no dramatic shift for any individual parameter. However, when considered as a whole, these posterior means for the elements of $\underset{\sim}{\theta}$ are significantly affected by a change in α. This is evident through a comparison of the posterior means using $\alpha=0$ with those using $\alpha=.10$, since it is seen that all $p=17$ posterior means differ in these two cases by at least one unit in the first decimal place.

Table 3a gives posterior probabilities of each y_u being discrepant conditional on a fixed number of observations r being discrepant. (Hence, these probabilities do no depend on the choice of α.) For $r=1$ discrepant observation, y_{10} is seen to be approximately seven times as likely to be the bad observation as is y_9. Furthermore, for $r=2$, y_9 is approximately three times as likely to be one of the 30 good observations as it is to be one of the two bad observations, while y_{10} is approximately five times as likely to be one of the bad values as it is to be one of the good values.

---

*In carrying out these analyses as well as similar analyses later in this chapter, it was further assumed that no more than two outliers were present in the data. In view of the values of α considered, this assumption is not unreasonable.

TABLE 3. CONDITIONAL POSTERIOR PROBABILITY THAT y_u IS DISCREPANT (I.E., HAS VARIANCE INFLATED BY A FACTOR $k^2 = 25$) GIVEN THAT EXACTLY r OBSERVATIONS ARE DISCREPANT: JOHN'S 2^5 DATA, USING M_2.

$$q_{u|1} = \Pr(y_u \text{ bad} \mid \underset{\sim}{y},\ r = 1)$$

$$q_{u|2} = \Pr(y_u \text{ bad} \mid \underset{\sim}{y},\ r = 2) = \sum_{\substack{v=1 \\ v \neq u}}^{n} \Pr(y_u \text{ and } y_v \text{ bad} \mid \underset{\sim}{y},\ r = 2)$$

u	(a) Untransformed ($\lambda = 1$) $q_{u\|1} \times 100$	$q_{u\|2} \times 100$	(b) Transformed ($\lambda = 0$) $q_{u\|1} \times 100$	$q_{u\|2} \times 100$
1	.486	2.057	1.584	3.155
2	.564	2.028	1.575	3.234
3	.492	1.319	3.599	7.842
4	.489	1.763	2.148	4.242
5	.673	2.139	1.632	3.247
6	.537	1.937	1.336	2.730
7	.550	1.994	1.849	6.314
8	.550	1.840	2.133	4.356
9	10.206	25.770	2.869	6.190
10	70.115	83.332	19.826	33.425
11	.733	5.005	1.335	2.754
12	.838	2.535	1.353	2.902
13	.649	3.740	1.375	3.143
14	.502	7.398	1.369	3.227
15	1.140	3.744	2.739	5.412
16	.982	4.610	5.015	10.579
17	.472	2.019	1.357	2.929
18	.479	1.951	1.866	4.917
19	.481	1.750	1.793	3.510
20	.492	1.821	1.657	4.237
21	.537	2.661	2.030	4.262
22	.616	7.379	2.134	6.139
23	.982	2.949	18.825	31.423
24	.859	7.331	3.997	9.919
25	.499	2.123	1.335	2.767
26	.597	2.032	1.463	3.029
27	.564	1.863	1.356	2.879
28	.557	2.224	1.351	2.726
29	.838	2.647	2.187	4.408
30	[illegible]	[illegible]	2.097	5.432
31	[illegible]	[illegible]	1.361	3.144
32	[illegible]	[illegible]	2.553	5.323

These findings seem to indicate that of these two observations, which are candidates for being discrepant based on the residual plot in Figure 1a , only y_{10} is indicated as truly being discrepant.

Whereas some investigators might be content with the above analysis of the data, as provided by the elaborated model which takes into account the possibility of bad data values, other investigators might wonder if there were alternate modeling options that should also be investigated. In particular, by noting that $y_{max}/y_{min} = 12.6/1.0$, the potential for an improved explanation of the data through transformation of the response variable y is indicated.

Thus, consider again the model (23) with the standard assumptions concerning the ε_u's. The diagnostic checking function (8) can be used to assess the need for transforming the y_u's. Informally, this involves observing the correlation between the residuals from fitting the y_u's to the model and the residuals from fitting the Y_u's to the model, with Y_u defined by (9). These two sets of residuals are given in Tables 1b and 1c , respectively, and they are plotted against each other in Figure 1b . Strong negative correlation is noticed (specifically, the calculated correlation is -.789), thus indicating the desirability to consider transformations of the form $y_u^{(\lambda)}$, given in (6) , with $\lambda<\lambda_0=1$.

It will therefore be beneficial to elaborate the basic model to provide for this needed transformation. This can be done using the method of Box and Cox (1964). (See also Box and Tiao, 1973).

Of particular interest will be the posterior distribution of λ based on a locally uniform prior for λ. (Equivalently, this will be the pseudo-likelihood function of λ.) Figure 2 shows the posterior distribution of λ not only for the model (23) (hereafter referred to as M_2) with $y_u^{(\lambda)}$ substituted for y_u, of course, but also for a constrained model M_1 which is strictly linear in the 6 input factors x_1-x_6 (i.e., M_1 excludes from M_2 the ten terms corresponding to all two factor interactions of factors 1-5). It is seen that both $p_u(\lambda|\underset{\sim}{y},M_2)$ and $p_u(\lambda|\underset{\sim}{y},M_1)$ strongly support the use of a logarithmic transformation (λ=0).

The constrained model M_1 is considered in addition to the original model M_2 because it is often the case that a suitable transformation will not only result in the assumptions on the ε_u's being more nearly satisfied but also will result in a simpler form of the response function being appropriate. The specific assessment of this aspect of transformation can be achieved through the consideration of Figure 3 , which is a plot* of $F(\lambda)$ versus λ, where $F(\lambda)$ is the F statistic appropriate for testing whether the 10 interaction terms of M_2 (that are not in M_1) can be excluded in fitting the transformed $y_u^{(\lambda)}$'s. [This statistic, which has as a predictive distribution the F distribution with 10 and 15 degrees of freedom when M_1 is true, is obtained from the analysis of variance table in

*There are additional plots that the investigator may wish to look at when considering the question of transformation; see, for example, Draper and Hunter (1969).

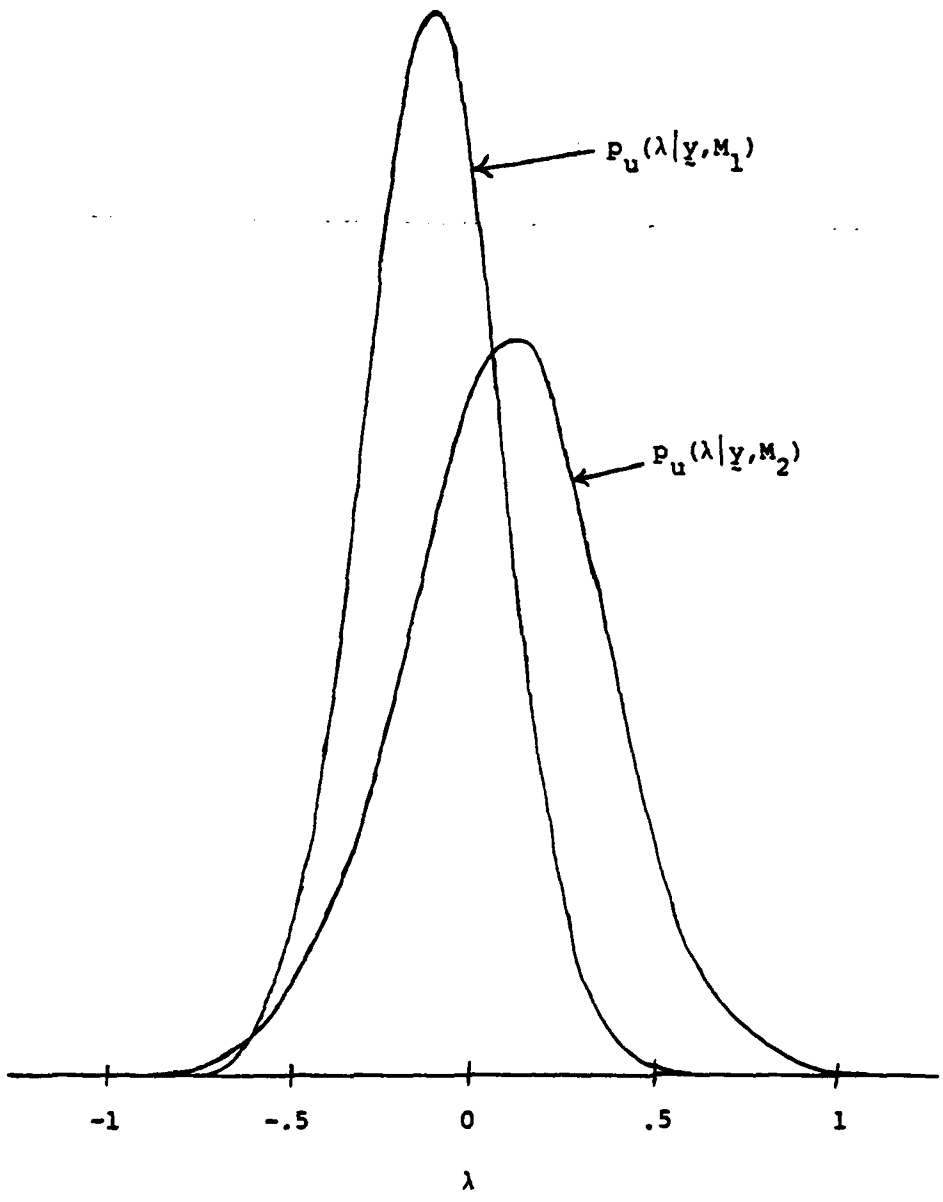


FIGURE 2. POSTERIOR DISTRIBUTIONS OF λ FOR DIFFERENT MODELS FOR JOHN'S 2^5 DATA

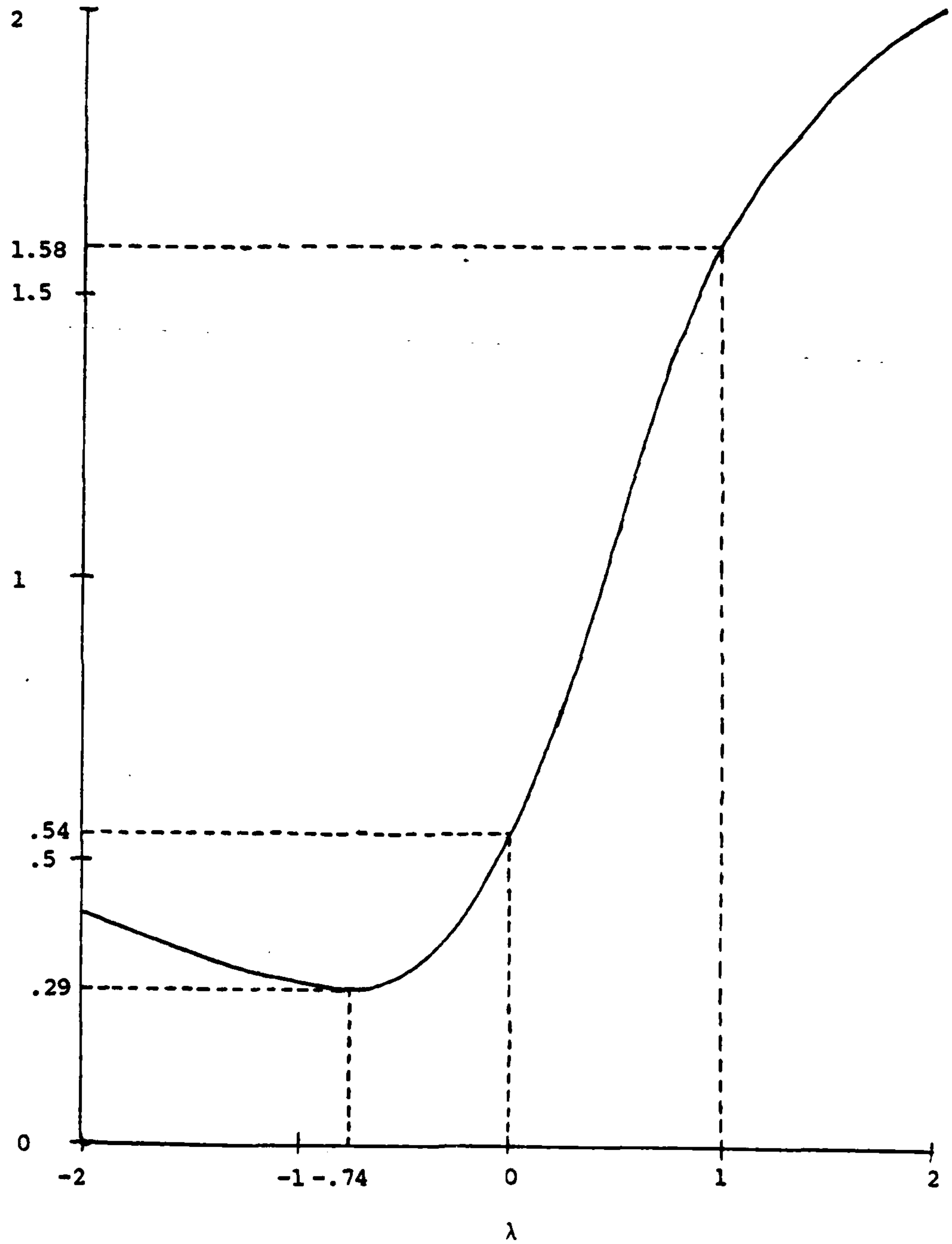


FIGURE 3. VALUES OF $F(\lambda)$ AS A FUNCTION OF λ FOR JOHN'S 2^5 DATA

the usual way as a ratio of mean squares. As an illustration, the analysis of variance tables are given in Tables 4a and 4b for the cases $\lambda=1$ (no transformation) and $\lambda=0$ (log transformation), respectively, from which it is seen that $F(1)=1.58$ and $F(0)=.54$.

The message to be gleaned from Figure 3 is that although there is no value of λ (at least in the range $-2\leq\lambda\leq 2$) for which the first-order model M_1 is judged inappropriate, all values of λ in the range $-2\leq\lambda\leq 0$ lead to a strikingly better first-order fit than is provided by the untransformed data ($\lambda=\lambda_0=1$).

In particular, consider the log transformation, as suggested by Figure 2. The transformed data, as well as the predicted values and residuals obtained from fitting to the unconstrained model M_2, are given in Table 1d. Figure 4a is a plot of these residuals versus their corresponding predicted values, and there is nothing about this plot that might indicate model inadequacy. Furthermore, when the logged data is fitted to the simplified first-order model M_1, the resulting plot of residuals versus predicted values, given in Figure 4b, also does not suggest any model inadequacy.

Further confirmation of the benefits derived from using the log transformation can be given by illustrating that it is no longer necessary to make allowances for outliers. This can be shown from both the diagnostic checking and the robustification points of view. With regard to diagnostic checking, not only does the residual plot in Figure 4a fail to indicate any spurious observations, but also the

TABLE 4. ANALYSIS OF VARIANCE FOR JOHN'S 2^5 DATA.

Source	(a) Untransformed ($\lambda = 1$)				(b) Transformed ($\lambda = 0$)			
	df	SS	MS	MS ratio	df	SS	MS	MS ratio
Main effects (Factors 1-6)	6	275.92	45.99	17.69	6	17.39	2.90	22.47(20.86)
Interactions (Factors 1-5)	10	41.06	4.11	1.58	10	.70	.07	.54(.50)
Residual	15	38.99	2.60	-	15(14)	1.94	.13(.14)	-
Total (corrected)	31	355.97	-	-	31	20.03	-	-

FIGURE 4. PLOTS OF RESIDUALS VS. PREDICTED VALUES FOR JOHN'S 2^5 DATA (USING LOG TRANSFORMATION).

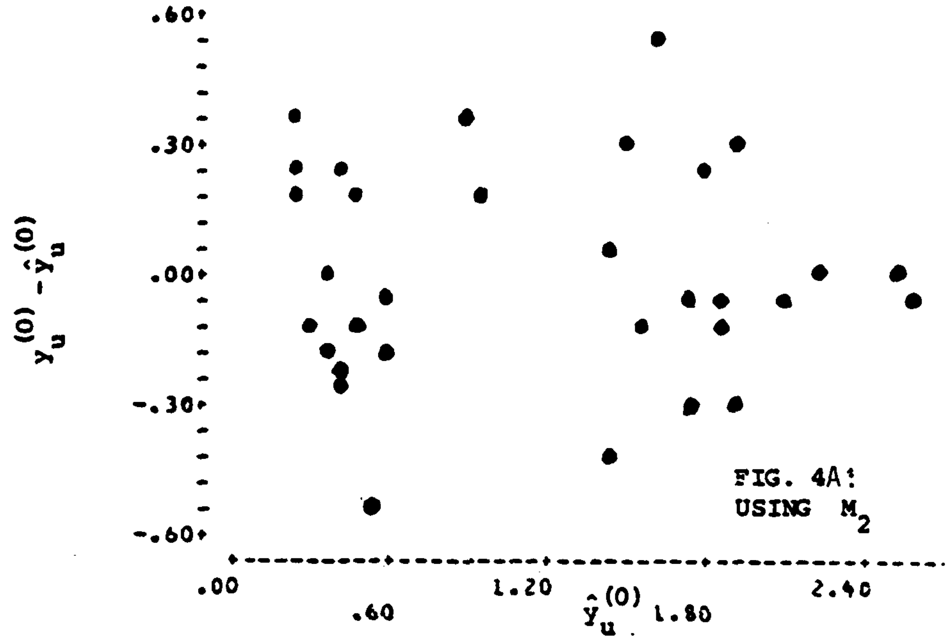


FIG. 4A: USING M_2

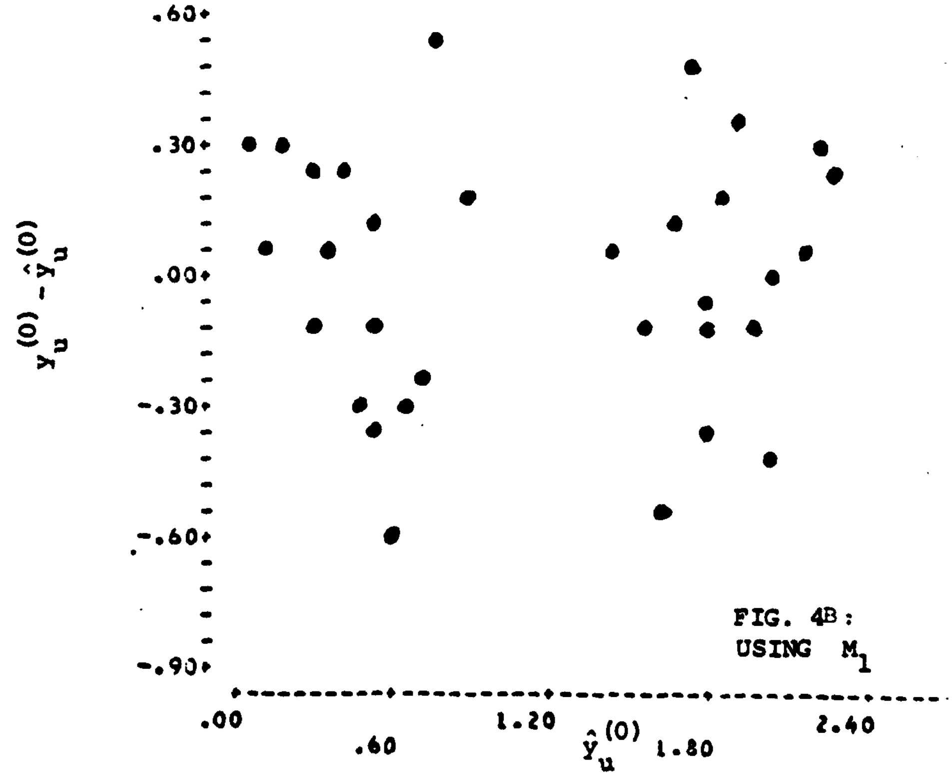


FIG. 4B: USING M_1

diagnostic check (11) is given by $g_{\alpha}(y) = n(h_1-1) = 32(.49-1) = -16.3$, and hence the negative value for this checking function does not discredit an assumption that $\alpha=0$. With regard to robustification, consideration of Tables 5 and 3b (which parallel Tables 2 and 3a for the untransformed data) shows that an assumption of $\alpha=0$ is not unjustified. Specifically, Table 5 gives the posterior means and the posterior standard deviations for the parameters of M_2 for various fixed values of α when fitting the logged data, and these posterior means as a whole are seen to exhibit somewhat more stable behavior throughout the range of α considered than was exhibited in Table 2 for the untransformed data. Furthermore, the posterior probabilities for any given y_u being discrepant conditional on a fixed number of outliers are given in Table 3b, and although the probabilities corresponding to y_{10} and y_{23} are somewhat larger than the other probabilities in both the $r=1$ and the $r=2$ cases, there is no compelling evidence that either of these observations are outliers.

We summarize the results of our investigation of John's data as follows:

(i) The application of various diagnostic checks indicates that a standard analysis of these data (without making allowances for model departures such as the presence of outliers or the need for transformation) is inappropriate.

(ii) Model robustification with respect to the possibility of outliers in the untransformed data shows the presence of exactly one discrepant value. Recalling that the residual

TABLE 5. POSTERIOR MEANS AND STD. DEVIATIONS (IN PARENTHESES) FOR θ's OF M_2 FOR JOHN'S 2^5 DATA (USING LOG TRANSFORMATION) FOR VARIOUS VALUES OF α WITH $k = 5$.

α	0	.001	.005	.010	.025	.050	.100
θ_0	1.170 (.068)	1.170 (.068)	1.170 (.069)	1.170 (.068)	1.170 (.069)	1.170 (.069)	1.170 (.069)
θ_1	-.167 (.068)	-.167 (.068)	-.168 (.068)	-.169 (.065)	-.172 (.069)	-.176 (.069)	-.180 (.068)
θ_2	.072 (.068)	.072 (.068)	.073 (.068)	.074 (.068)	.077 (.069)	.081 (.068)	.084 (.068)
θ_3	-.136 (.068)	-.136 (.068)	-.135 (.068)	-.133 (.068)	-.130 (.069)	-.126 (.068)	-.122 (.068)
θ_4	.699 (.068)	.698 (.069)	.697 (.068)	.696 (.068)	.694 (.069)	.691 (.068)	.687 (.068)
θ_5	.061 (.068)	.061 (.068)	.062 (.068)	.063 (.068)	.066 (.068)	.069 (.068)	.072 (.068)
θ_{12}	-.104 (.068)	-.104 (.068)	-.104 (.068)	-.104 (.068)	-.104 (.069)	-.103 (.069)	-.103 (.069)
θ_{13}	.047 (.068)	.047 (.068)	.047 (.068)	.047 (.068)	.047 (.069)	.047 (.069)	.047 (.069)
θ_{14}	-.047 (.068)	-.047 (.068)	-.047 (.068)	-.047 (.068)	-.047 (.069)	-.047 (.069)	-.046 (.069)
θ_{15}	-.003 (.068)	-.003 (.068)	-.003 (.068)	-.003 (.068)	-.004 (.069)	-.004 (.069)	-.004 (.069)
θ_{23}	-.037 (.068)	-.037 (.068)	-.037 (.068)	-.037 (.068)	-.037 (.069)	-.037 (.069)	-.037 (.070)
θ_{24}	.014 (.068)	.014 (.068)	.014 (.068)	.014 (.068)	.015 (.069)	.015 (.069)	.016 (.070)
θ_{25}	-.040 (.068)	-.040 (.068)	-.040 (.068)	-.040 (.068)	-.040 (.069)	-.041 (.069)	-.042 (.069)
θ_{34}	-.043 (.068)	-.043 (.068)	-.043 (.068)	-.042 (.068)	-.042 (.069)	-.042 (.069)	-.042 (.070)
θ_{35}	-.000 (.068)	-.000 (.068)	-.001 (.068)	-.001 (.068)	-.001 (.069)	-.002 (.069)	-.003 (.069)
θ_{45}	.039 (.068)	.039 (.068)	.039 (.068)	.039 (.068)	.039 (.069)	.039 (.069)	.039 (.069)
θ_6	-.019 (.068)	-.019 (.068)	-.018 (.068)	-.017 (.068)	-.014 (.069)	-.011 (.068)	-.007 (.068)

plot in Figure 1a (which was considered during the diagnostic checking of the standard model) suggested that as many as two observations may be discrepant, we thus see that the situation is clarified through the robustification of this model.

(iii) An improved explanation of the data can be obtained by considering model robustification with respect to the possible need for transformation of the response variable (rather than with respect to the possibility of outliers), and this leads to an analysis based on the logarithmically transformed data.

5. <u>A diagnostic check for transformation of the predictor variables</u>.

In the example just discussed, the diagnostic check $g_{\lambda}(\underset{\sim}{y})$ indicated the need to consider a transformation of the response variables y in order to obtain an adequate representation of the relationship between y and the input variables, given in coded form by the x_i's.

In situations where the k input variables under consideration — $\underset{\sim}{\xi}=(\xi_1,\ldots,\xi_k)'$ in uncoded form — are all quantitative (i.e., can take on any value over some continuous range), it may be the case that a better empirical relationship between $\underset{\sim}{\xi}$ and y can be developed by dealing with transformations of some or all of the elements of $\underset{\sim}{\xi}$ rather than with $\underset{\sim}{\xi}$ itself.

In particular, consider the normal linear model for which $p(\underset{\sim}{y}|\underset{\sim}{\theta},\sigma^2,\underset{\sim}{\lambda})$ is $N_n(X^{(\underset{\sim}{\lambda})}\underset{\sim}{\theta},\sigma^2 I)$, so that for the u^{th} observation $p(y_u|\underset{\sim}{\theta},\sigma^2,\underset{\sim}{\lambda})$ is $N(\underset{\sim}{x}_u^{(\underset{\sim}{\lambda})'}\underset{\sim}{\theta},\sigma^2)$, with $\underset{\sim}{x}_u^{(\underset{\sim}{\lambda})'}$ corresponding to the u^{th} row of $X^{(\underset{\sim}{\lambda})}$. The vector $\underset{\sim}{\lambda} = (\lambda_1,\ldots,\lambda_k)'$ denotes that each $\underset{\sim}{x}_u^{(\underset{\sim}{\lambda})'}$ depends on the u^{th} set of inputs $\underset{\sim}{\xi}_u = (\xi_{1u},\ldots,\xi_{ku})'$ only through the transformed set of inputs $\underset{\sim}{\xi}_u^{(\underset{\sim}{\lambda})} = (\xi_{1u}^{(\lambda_1)},\ldots,\xi_{ku}^{(\lambda_k)})'$ where

$$\xi_{iu}^{(\lambda_i)} = \begin{cases} \xi_{iu}^{\lambda_i} & \lambda_i \neq 0, \\ \ln \xi_{iu} & \lambda_i = 0. \end{cases} \tag{24}$$

Note that $\underset{\sim}{\xi}_u = \underset{\sim}{\xi}_u^{(\underset{\sim}{\lambda}_0)}$, where $\underset{\sim}{\lambda}_0 = \underset{\sim}{1}$. Thus, for example, the investigator may wish to fit a low order polynomial model in the untransformed predictor variables $\underset{\sim}{\xi}$ to the data, when it may actually be more appropriate to be using the same order polynomial (or perhaps even a polynomial of lower order) expressed in terms of some suitably transformed predictor variables $\underset{\sim}{\xi}^{(\underset{\sim}{\lambda})}$. For this situation it would be useful to have available a diagnostic check which could be used to assess the appropriateness of using $\underset{\sim}{\lambda} = \underset{\sim}{\lambda}_0 = \underset{\sim}{1}$. Specifically, the particular checking function suggested by (1) can be considered for each λ_i. The derivation of this predictive check is now given.

We assume that, for any specified $\underset{\sim}{\lambda}$, $p(\underset{\sim}{\theta},\sigma^2|\underset{\sim}{\lambda})$ is locally uniform in $\theta_1,\ldots,\theta_p$ and $\ln \sigma$. However, since prior independence between $\underset{\sim}{\theta}$ and $\underset{\sim}{\lambda}$ is clearly an inappropriate assumption, then following the argument of Pallesen (1977, Section 2.3.1), we employ the prior

$$p(\underset{\sim}{\theta},\sigma^2|\underset{\sim}{\lambda}) \propto |X^{(\underset{\sim}{\lambda})'}X^{(\underset{\sim}{\lambda})}|^{\frac{1}{2}}\frac{1}{\sigma^2} . \tag{25}$$

This combines with

$$p(\underset{\sim}{y}|\underset{\sim}{\theta},\sigma^2,\underset{\sim}{\lambda}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}(\underset{\sim}{y}-X^{(\underset{\sim}{\lambda})}\underset{\sim}{\theta})'(\underset{\sim}{y}-X^{(\underset{\sim}{\lambda})}\underset{\sim}{\theta})\} \tag{26}$$

to give $p(\underset{\sim}{y},\underset{\sim}{\theta},\sigma^2|\underset{\sim}{\lambda})$ which, upon integrating out $\underset{\sim}{\theta}$ and σ^2, yields the conditional predictive distribution

$$p(\underset{\sim}{y}|\underset{\sim}{\lambda}) \propto (\nu s^2_{\underset{\sim}{\lambda}})^{-\frac{\nu}{2}} , \tag{27}$$

where

$$\left.\begin{aligned} \nu s^2_{\underset{\sim}{\lambda}} &= (\underset{\sim}{y}-X^{(\underset{\sim}{\lambda})}\hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}})'(\underset{\sim}{y}-X^{(\underset{\sim}{\lambda})}\hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}}) = \sum_{u=1}^{n}(y_u-\underset{\sim}{x}_u{}^{(\underset{\sim}{\lambda})'}\hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}})^2, \\ \hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}} &= [X^{(\underset{\sim}{\lambda})'}X^{(\underset{\sim}{\lambda})}]^{-1}X^{(\underset{\sim}{\lambda})'}\underset{\sim}{y}, \quad \nu = n-p. \end{aligned}\right\} \tag{28}$$

Hence,

$$\frac{\partial}{\partial\lambda_i} \ln p(\underset{\sim}{y}|\underset{\sim}{\lambda}) = -\frac{\nu}{2}[(\frac{1}{\nu s^2_{\underset{\sim}{\lambda}}})(\frac{\partial}{\partial\lambda_i}\nu s^2_{\underset{\sim}{\lambda}})]$$

$$= -\frac{1}{2s^2_{\underset{\sim}{\lambda}}}\sum_{u=1}^{n}[\frac{\partial}{\partial\lambda_i}(y_u-\underset{\sim}{x}_u{}^{(\underset{\sim}{\lambda})'}\hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}})^2]$$

$$= -\frac{1}{2s^2_{\underset{\sim}{\lambda}}}\sum_{u=1}^{n}\{-2(y_u-\underset{\sim}{x}_u{}^{(\underset{\sim}{\lambda})'}\hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}})[\frac{\partial}{\partial\lambda_i}(\underset{\sim}{x}_u{}^{(\underset{\sim}{\lambda})'}\hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}})]\}$$

$$= \frac{1}{s^2_{\underset{\sim}{\lambda}}}\sum_{u=1}^{n}\{(y_u-\underset{\sim}{x}_u{}^{(\underset{\sim}{\lambda})'}\hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}})[(\frac{\partial}{\partial\lambda_i}\underset{\sim}{x}_u{}^{(\underset{\sim}{\lambda})})'\hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}}+\underset{\sim}{x}_u{}^{(\underset{\sim}{\lambda})'}(\frac{\partial}{\partial\lambda_i}\hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}})]\} . \tag{29}$$

Thus, evaluation of this expression at $\underset{\sim}{\lambda} = \underset{\sim}{\lambda}_0 = \underset{\sim}{1}$ gives

$$g_{\lambda_i}(\underset{\sim}{y}) = g^{(BT)}_{\lambda_i}(\underset{\sim}{y}) + g^{(P)}_{\lambda_i}(\underset{\sim}{y}), \tag{30}$$

with

$$g^{(BT)}_{\lambda_i}(\underset{\sim}{y}) = \frac{1}{s^2}\sum_{u=1}^{n}\{(y_u - \underset{\sim}{x}_u'\hat{\underset{\sim}{\theta}})[\frac{\partial}{\partial\lambda_i}(\underset{\sim}{x}_u^{(\underset{\sim}{\lambda})'}\hat{\underset{\sim}{\theta}})]\Big|_{\underset{\sim}{\lambda}=\underset{\sim}{1}}\}$$

$$= \frac{\underset{\sim}{y}'R\underset{\sim}{z}_i^{(BT)}}{s^2} \tag{31}$$

and

$$g^{(P)}_{\lambda_i}(\underset{\sim}{y}) = \frac{1}{s^2}\sum_{u=1}^{n}\{(y_u - \underset{\sim}{x}_u'\hat{\underset{\sim}{\theta}})[\frac{\partial}{\partial\lambda_i}(\underset{\sim}{x}_u'\hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}})]\Big|_{\underset{\sim}{\lambda}=\underset{\sim}{1}}\}$$

$$= \frac{\underset{\sim}{y}'R\underset{\sim}{z}_i^{(P)}}{s^2} \quad ; \tag{32}$$

where $\underset{\sim}{z}_i^{(BT)} = (z_{i1}^{(BT)},\ldots,z_{in}^{(BT)})'$ and $\underset{\sim}{z}_i^{(P)} = (z_{i1}^{(P)},\ldots,z_{in}^{(P)})'$ with

$$z_{iu}^{(BT)} = [\frac{\partial}{\partial\lambda_i}(\underset{\sim}{x}_u^{(\underset{\sim}{\lambda})'}\hat{\underset{\sim}{\theta}})]\Big|_{\underset{\sim}{\lambda}=\underset{\sim}{1}} \tag{33}$$

and

$$z_{iu}^{(P)} = [\frac{\partial}{\partial\lambda_i}(\underset{\sim}{x}_u'\hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}})]\Big|_{\underset{\sim}{\lambda}=\underset{\sim}{1}}, \tag{34}$$

and where

$$\left.\begin{array}{l} \underset{\sim}{x}_u = \underset{\sim}{x}_u^{(\underset{\sim}{\lambda}_0)},\ R = X(X'X)^{-1}X',\ X = X^{(\underset{\sim}{\lambda}_0)}, \\ \\ \hat{\underset{\sim}{\theta}} = \hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}_0},\ s^2 = s^2_{\underset{\sim}{\lambda}_0} \end{array}\right\} \tag{35}$$

denote the appropriate quantities for an analysis based on assuming $\underset{\sim}{\lambda} = \underset{\sim}{\lambda}_0 = \underset{\sim}{1}$.

Writing

$$\underset{\sim}{z}_i = \underset{\sim}{z}_i^{(BT)} + \underset{\sim}{z}_i^{(P)}, \tag{36}$$

the overall predictive check for the i^{th} predictor variable ξ_i is thus

$$g_{\lambda_i}(\underset{\sim}{y}) = \frac{\underset{\sim}{y}'R\underset{\sim}{z}_i}{s^2}, \tag{37}$$

and it has the interpretation of seeking a correlation between the residuals from fitting $\underset{\sim}{y}$ to $N_n(X\underset{\sim}{\theta},\sigma^2 I)$ and the residuals from fitting $\underset{\sim}{z}_i$ to the same model. Strong positive correlation will suggest that values of $\lambda_i>1$ may be more appropriate, while strong negative correlation will suggest that values of $\lambda_i<1$ may need to be considered.

In considering the two components which combine to form $g_{\lambda_i}(\underset{\sim}{y})$ according to (30), it is noted in particular that the first component $g_{\lambda_i}^{(BT)}(\underset{\sim}{y})$ can be interpreted in the context of the iterative procedure for the joint estimation of $\underset{\sim}{\theta}$ and $\underset{\sim}{\lambda}$ proposed by Box and Tidwell (1962). (See also Box and Draper, 1979.) The steps involved in the first iteration of this procedure are as follows (using the notation defined above):

(i) Fit the model $N_n(X\underset{\sim}{\theta},\sigma^2 I)$ to the data $\underset{\sim}{y}$.

(ii) Use $\hat{\underset{\sim}{\theta}}$ from (i) to obtain the matrix $Z^{(BT)}$ having entries $z_{iu}^{(BT)}$ given by (33).

(iii) Fit the model $N_n(X\underset{\sim}{\theta}+Z^{(BT)}(\underset{\sim}{\lambda}-\underset{\sim}{1}),\sigma^2 I)$.

In the particular case where transformation of only the i^{th} predictor variable ξ_i is contemplated, step (iii) above reduces to fitting the model $N_n(X\underset{\sim}{\theta}+\underset{\sim}{z}_i^{(BT)}(\lambda_i-1),\sigma^2 I)$, and the relevance of $g_{\lambda_i}(\underset{\sim}{y})$ as a diagnostic checking function in this context is readily seen.

It is interesting to note that the estimated $\hat{\underset{\sim}{\theta}}$, which is calculated using an assumption that $\underset{\sim}{\lambda} = \underset{\sim}{\lambda}_0 = \underset{\sim}{1}$ is used in step (ii) above as a matter of convenience, according to the authors; whereas explicit consideration of the nature of $\hat{\underset{\sim}{\theta}}_{\underset{\sim}{\lambda}}$ with respect to $\underset{\sim}{\lambda}$ is provided through the presence of the second term, $g_{\lambda_i}^{(P)}(\underset{\sim}{y})$, in the function $g_{\lambda_i}(\underset{\sim}{y})$. It can thus be expected that diagnostic checking based on $g_{\lambda_i}(\underset{\sim}{y})$ will be more sensitive to the need for a power transformation of ξ_i than diagnostic checking based on the results of the first iteration of the Box-Tidwell procedure (and hence based on $g_{\lambda_i}^{(BT)}(\underset{\sim}{y})$).

6. Summary

The question of what tentatively to include in a model and what tentatively to omit from a model merits careful consideration by the investigator. The answers, of course, will depend upon the

nature of the phenomenon that the investigator is attempting to model. What is clear, however, is that there is a need for the investigator to have for his use not only (i) well developed techniques which allow for model elaboration with respect to those aspects of model departure that he fears the most and/or are the most likely to occur, but also (ii) suitable diagnostic checks which allow for model criticism with respect to those aspects not explicitly provided for in the model. The techniques developed and discussed in this paper are of use in filling these needs.

7. Two final examples.

In the introduction of this paper, reference was made to the suggestion of Chen and Box (1979a,b) that a contaminated normal distribution will often be a reasonable choice of an error distribution when modeling data from a well planned experiment. To illustrate the practical application of this suggestion, a side-by-side comparison of two data sets analyzed from this viewpoint is presented in this section.

Both sets of n=27 observations were from experiments using the same design. The design, displayed in Table 6a, can be used in fitting a second degree polynomial in four factors, denoted by

$$y_u=\theta_0+\sum_{i=1}^{4}\theta_i x_{iu}+\sum_{i=1}^{4}\theta_{ii}x_{iu}^2+\sum_{i=1}^{3}\sum_{j=i+1}^{4}\theta_{ij}x_{iu}x_{ju}+\theta_L x_{Lu}+\theta_Q x_{Qu}+\varepsilon_u;u=1,\ldots,27, \quad (38)$$

TABLE 6. RESULTS FROM TWO EXPERIMENTS USING FOUR FACTOR BOX-BEHNKEN DESIGN.

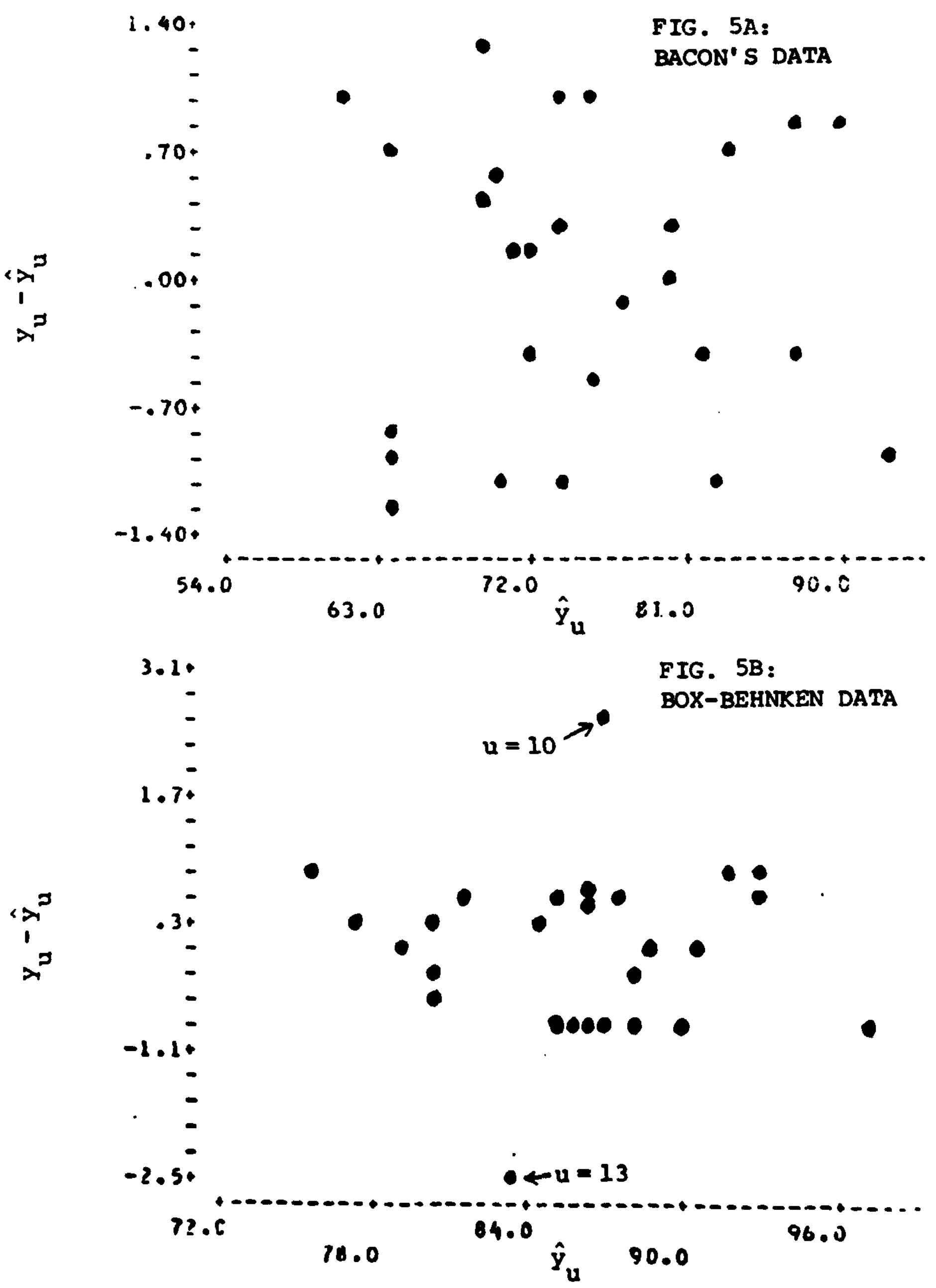
	(a) The design						(b) Bacon's data			(c) Box-Behnken data		
u	x_{1u}	x_{2u}	x_{3u}	x_{4u}	x_{Lu}	x_{Qu}	y_u	$\hat{y}_u$	$y_u - \hat{y}_u$	y_u	$\hat{y}_u$	$y_u - \hat{y}_u$
1	1	-1	-1	0	-1	1	84.5	83.74	.76	84.7	85.53	-.83
2	1	1	-1	0	-1	1	62.9	63.81	-.91	93.3	92.75	.55
3	1	-1	1	0	-1	1	90.7	89.88	.82	84.2	84.97	-.77
4	1	1	1	0	-1	1	63.2	64.04	-.84	86.1	85.48	.62
5	1	0	0	-1	-1	1	70.9	70.32	.58	85.7	86.40	-.70
6	1	0	0	1	-1	1	69.2	70.32	-1.12	96.4	97.17	-.77
7	1	0	0	-1	-1	1	80.1	79.80	.30	88.1	87.55	.55
8	1	0	0	1	-1	1	79.8	79.80	-.00	81.8	81.32	.48
9	1	0	0	0	-1	1	75.1	74.18	.92	93.8	92.93	.87
10	1	-1	0	0	0	-2	81.8	82.89	-1.09	89.4	86.79	2.61
11	1	1	0	0	0	-2	61.8	60.76	1.04	88.7	88.76	-.06
12	1	-1	0	0	0	-2	92.2	93.13	-.93	77.8	77.54	.26
13	1	1	0	0	0	-2	70.7	69.49	1.21	80.9	83.31	-2.41
14	1	0	-1	-1	0	-2	72.4	72.27	.13	80.9	80.68	.22
15	1	0	1	-1	0	-2	76.4	75.45	.95	79.8	80.11	-.31
16	1	0	-1	1	0	-2	71.9	72.27	-.37	86.8	86.29	.51
17	1	0	1	1	0	-2	74.9	75.45	-.55	79.0	79.03	-.03
18	1	0	0	0	0	-2	74.5	74.18	.32	87.3	88.10	-.80
19	1	0	-1	0	1	1	69.4	68.91	.49	86.1	86.87	-.77
20	1	0	1	0	1	1	71.2	71.09	.11	87.9	88.21	-.31
21	1	0	-1	0	1	1	77.3	77.39	-.09	85.1	84.77	.33
22	1	0	1	0	1	1	81.1	81.58	-.48	76.4	75.61	.79
23	1	-1	0	-1	1	1	88.0	87.13	.87	79.7	80.22	-.52
24	1	1	0	-1	1	1	63.0	64.24	-1.24	92.5	91.74	.76
25	1	-1	0	1	1	1	86.7	87.13	-.43	89.4	90.14	-.74
26	1	1	0	1	1	1	65.0	64.24	.76	86.9	86.36	.54
27	1	0	0	0	1	1	73.0	74.18	-1.18	90.7	90.77	-.07

where, as indicated above, the ε_u's will be assumed to follow a contaminated normal distribution. (Specifically, the type of contamination used in the Section 4 example will also be used here, so that each observation has a probability α of being contaminated in that it has a standard deviation which is k=5 times as large as the standard deviation of an uncontaminated observation.) This design can be divided into three orthogonal blocks, which explains the presence of the linear and quadratic blocking variables x_{Lu} and x_{Qu} in (38).

The two sets of data considered are those of Bacon (1970) and of Box and Behnken (1960). (The latter reference is, in fact, the paper in which the design in Table 6a was introduced.) These data are given in Tables 6b and 6c, along with the corresponding predicted values and residuals from fitting (38) to each set of data under the assumption of no contamination (α=0). Actually, for the Bacon data, a simpler model than (38) was found to be adequate, in that terms involving x_3 and the two blocking terms were omitted from the model.

Plots of the residuals versus the predicted values are shown in Figure 5. For the Bacon data (Figure 5a) no unusual behavior is noticed. However, for the Box-Behnken data (Figure 5b), the possibility of two bad values (y_{10} and y_{13}) is suggested. It is of interest to explore how the analysis of each of these sets

FIGURE 5. PLOTS OF RESIDUALS VS. PREDICTED VALUES FOR DATA FROM BOX-BEHNKEN DESIGN.



of data is affected when the possibility of discrepant observations is allowed for by choosing $\alpha \neq 0$ and using the techniques developed by Box and Tiao (1968).

Considering the Bacon data first, the posterior means and standard deviations for the p=10 parameters used in the model are shown in Table 7 for various choices of α, and these quantities are seen to be remarkably stable over the range of α studied. Furthermore, the posterior probabilities for any specified y_u being discrepant conditional on a fixed number of outliers r, which are given in Table 8a for the cases r=1 and r=2, do not indicate the presence of any bad observations.

Thus the conclusion arrived at by visual inspection of the residual plot in Figure 5a — namely, that there are no discrepant observations in Bacon's data — is further supported, from the robustification point of view, by an analysis using a model which takes into account the possibility of bad observations.

Of course, other model criticism techniques should be employed to explore other areas in which the model may be improved upon. For example, consider the analysis of variance displayed in Table 9a. In this table, the residual sum of squares obtained from fitting to the data a full second order polynomial model in all four factors (including x_3) is partitioned in such a way so as to provide diagnostic checks not only for the need to account for block to block variation, but also for the possibility that a third or higher order polynomial model might be needed to adequately approximate the underlying response surface.

Specifically, letting θ_{iii}, θ_{ijj}, and θ_{ijl} denote coefficients of x_i^3, $x_i x_j^2$, and $x_i x_j x_l$, respectively, in a third order polynomial representation; we have the following properties of this design:

(i) The design does not allow for the separate estimation of θ_i and θ_{iii}, since $x_i = x_i^3$ for all design points.

(ii) It does not allow for the estimation of θ_{ijl}, since $x_i x_j x_l = 0$ for all design points.

(iii) It does not allow for the separate estimation of the θ_{ijj}'s, but does allow for the estimation of certain contrasts of the θ_{ijj}'s <u>independent</u> of the estimation of coefficients in the second order model.

TABLE 7 POSTERIOR MEANS AND STD. DEVIATIONS (IN PARENTHESES) FOR θ's OF MODEL FOR BACON'S DATA, FOR VARIOUS VALUES OF α WITH $k = 5$.

α	0	.001	.005	.01	.025	.05	.10
θ_0	74.18 (.45)	74.18 (.45)	74.18 (.45)	74.18 (.45)	74.18 (.45)	74.18 (.45)	74.19 (.44)
θ_1	-11.44 (.30)	-11.44 (.30)	-11.44 (.30)	-11.45 (.31)	-11.45 (.31)	-11.47 (.32)	-11.48 (.34)
θ_2	1.59 (.30)	1.59 (.30)	1.59 (.30)	1.59 (.30)	1.59 (.30)	1.59 (.30)	1.59 (.30)
θ_4	4.74 (.30)	4.74 (.30)	4.74 (.30)	4.74 (.30)	4.74 (.31)	4.73 (.31)	4.72 (.32)
θ_{11}	1.51 (.43)	1.51 (.43)	1.51 (.43)	1.51 (.43)	1.50 (.43)	1.50 (.44)	1.50 (.45)
θ_{22}	-.32 (.43)	-.32 (.43)	-.32 (.43)	-.32 (.43)	-.32 (.43)	-.32 (.43)	-.32 (.43)
θ_{44}	.88 (.43)	.88 (.43)	.88 (.43)	.88 (.43)	.88 (.43)	.87 (.44)	.87 (.44)
θ_{12}	-1.48 (.52)	-1.48 (.52)	-1.48 (.52)	-1.48 (.53)	-1.48 (.53)	-1.48 (.54)	-1.48 (.56)
θ_{14}	-.38 (.52)	-.38 (.52)	-.38 (.53)	-.38 (.54)	-.38 (.57)	-.39 (.61)	-.39 (.68)
θ_{24}	.50 (.52)	.50 (.52)	.50 (.52)	.50 (.52)	.50 (.52)	.50 (.52)	.50 (.51)

TABLE 8. CONDITIONAL POSTERIOR PROBABILITY THAT y_u IS DISCREPANT (I.E., HAS VARIANCE INFLATED BY A FACTOR $k^2 = 25$) GIVEN THAT EXACTLY r OBSERVATIONS ARE DISCREPANT: DATA FROM FOUR FACTOR BOX-BEHNKEN DESIGN.

$$q_{u|1} = \Pr(y_u \text{bad} | \underset{\sim}{y}, r = 1)$$

$$q_{u|2} = \Pr(y_u \text{bad} | \underset{\sim}{y}, r = 2) = \sum_{\substack{v=1 \\ v \neq u}}^{n} \Pr(y_u \text{ and } y_v \text{bad} | \underset{\sim}{y}, r = 2)$$

	(a) Bacon's data		(b) Box-Behnken data	
u	$q_{u\|1} \times 100$	$q_{u\|2} \times 100$	$q_{u\|1} \times 100$	$q_{u\|2} \times 100$
1	4.069	7.666	.018	3.734
2	5.594	10.261	.014	2.956
3	4.623	8.609	.016	3.346
4	4.838	9.355	.014	3.032
5	1.955	3.878	.015	3.110
6	3.605	8.287	.016	3.346
7	1.665	3.326	.014	2.958
8	1.578	3.060	.013	2.955
9	2.647	5.212	.012	6.418
10	8.976	18.335	93.898	92.527
11	7.684	16.223	.011	3.046
12	5.813	13.433	.012	3.354
13	12.671	23.205	5.764	16.549
14	1.595	3.052	.012	3.225
15	2.829	5.730	.012	4.386
16	1.720	3.372	.013	6.416
17	1.915	3.827	.011	2.972
18	1.653	3.145	.011	4.658
19	2.753	5.5[illegible]	.010	3.375
20	2.126	4.17[illegible]	.012	4.121
21	2.121	4.134	.012	3.496
22	[illegible].719	5.351	.017	3.835
23	2.582	5.[illegible]	.013	3.029
24	4.435	9.217	.016	3.596
25	1.771	3.5[illegible]	.016	3.236
26	[illegible].277	5.1[illegible]	.013	2.956
27	[illegible].775	7.9[illegible]	.009	2.354

TABLE 9. ANALYSES OF VARIANCE FOR DATA GENERATED BY THE FOUR FACTOR BOX-BEHNKEN DESIGN.

				(a) Bacon's data		(b) Box-Behnken data	
Source			df	SS	MS	SS	MS
First order terms			4	1872.19	468.05	268.36	67.09
Second order terms			10	33.09	3.31	294.92	29.49
Residual	Blocks		2	.24	.12	105.53	52.77
	Third order terms	L_1	2	7.55	3.77	10.67	5.34
		L_2	2	.25	.13	.51	.25
		L_3	2	1.22	.61	.13	.07
		L_4	2	.54	.27	7.30	3.65
	Higher order terms		2	2.14	1.07	1.57	.79
Total (corrected)			26	1917.22	-	689.99	-

With regard to (iii), each sum of squares denoted by L_i in Table 9a is associated with all linear contrasts of the θ_{ijj} estimates for $j \neq i$. The remaining sum of squares at the bottom of the table is associated with coefficients of order higher than three. (A similar type of partitioning has been previously explored by Draper and Herzberg, 1971, for the central composite response surface designs; for some related work involving single degree of freedom contrasts, see Box, Hunter and Hunter, 1978, and Box and Draper, 1979.) The somewhat large sum of squares given for L_1 in the table suggests possible lack of fit of a second order polynomial due to contrasts of θ_{122}, θ_{133}, and θ_{144} of significant non-zero magnitude. Practically speaking, this suggests that a closer approximation to the true surface may be gained by transforming one or more of the four factors before fitting the second order polynomial (Box and Draper, 1979), but this possibility will not be explored in further detail here.

Turning now to the Box-Behnken data (Table 6c), Table 10 gives the posterior means and standard deviations for the p = 17 parameters in the model (38). It is seen that many of these quantities undergo a noticeable change when the contamination frequency parameter α is allowed to assume non-zero values. The most dramatic of the changes in the expectations is observed in the posterior mean for θ_{14}, where the contemplation of even a small non-zero value for α switches the sign of $E(\theta_{14}|\underset{\sim}{y},\alpha)$. Moreover, for all parameters except θ_{14}, the

TABLE 10. POSTERIOR MEANS AND STD. DEVIATIONS (IN PARENTHESES) FOR θ's OF MODEL FOR THE BOX-BEHNKEN DATA, FOR VARIOUS VALUES OF α WITH $k=5$.

α	0	.001	.005	.01	.025	.05	.10
θ_0	90.60 (.94)	90.60 (.56)	90.60 (.45)	90.60 (.42)	90.60 (.41)	90.60 (.41)	90.60 (.41)
θ_1	1.93 (.47)	2.35 (.39)	2.46 (.28)	2.48 (.25)	2.49 (.23)	2.49 (.22)	2.49 (.22)
θ_2	-1.96 (.47)	-1.96 (.29)	-1.96 (.22)	-1.96 (.21)	-1.96 (.20)	-1.96 (.20)	-1.96 (.20)
θ_3	1.13 (.47)	1.13 (.29)	1.13 (.22)	1.13 (.21)	1.13 (.20)	1.13 (.20)	1.13 (.20)
θ_4	-3.68 (.47)	-3.26 (.39)	-3.15 (.28)	-3.13 (.25)	-3.12 (.23)	-3.12 (.22)	-3.12 (.22)
θ_{11}	-1.42 (.70)	-1.79 (.54)	-1.88 (.44)	-1.90 (.42)	-1.90 (.41)	-1.90 (.41)	-1.89 (.42)
θ_{22}	-4.33 (.70)	-4.14 (.46)	-4.10 (.36)	-4.09 (.35)	-4.09 (.34)	-4.09 (.34)	-4.09 (.34)
θ_{33}	-2.24 (.70)	-2.05 (.46)	-2.01 (.36)	-2.00 (.35)	-2.00 (.34)	-2.00 (.34)	-2.01 (.34)
θ_{44}	-2.58 (.70)	-2.95 (.54)	-3.05 (.44)	-3.06 (.42)	-3.06 (.41)	-3.06 (.41)	-3.05 (.42)
θ_{12}	-1.67 (.81)	-1.67 (.50)	-1.67 (.39)	-1.67 (.36)	-1.67 (.35)	-1.67 (.35)	-1.67 (.34)
θ_{13}	-3.83 (.81)	-3.82 (.50)	-3.82 (.39)	-3.82 (.36)	-3.82 (.35)	-3.82 (.34)	-3.82 (.34)
θ_{14}	.95 (.81)	-.17 (1.03)	-.46 (.95)	-.49 (.93)	-.51 (.92)	-.50 (.94)	-.47 (.96)
θ_{23}	-1.67 (.81)	-1.67 (.50)	-1.67 (.39)	-1.67 (.36)	-1.67 (.35)	-1.67 (.35)	-1.67 (.35)
θ_{24}	-2.62 (.81)	-2.62 (.50)	-2.62 (.39)	-2.62 (.36)	-2.62 (.35)	-2.62 (.35)	-2.62 (.35)
θ_{34}	-4.25 (.81)	-4.25 (.50)	-4.25 (.39)	-4.25 (.36)	-4.25 (.35)	-4.25 (.34)	-4.25 (.34)
θ_L	-1.08 (.38)	-1.08 (.24)	-1.08 (.18)	-1.08 (.17)	-1.08 (.16)	-1.08 (.16)	-1.08 (.16)
θ_Q	1.25 (.22)	1.42 (.20)	1.46 (.17)	1.47 (.16)	1.47 (.16)	1.47 (.16)	1.46 (.16)

corresponding posterior standard deviations are about halved, which effectively increases the size and sensitivity of the experiment by a factor of four.

Note, however, that from Table 8b the posterior probabilities for y_u being bad, given that exactly r values are bad, suggest that there is only one bad observation and not two, as first suggested by the Figure 5b residual plot. Specifically, for r = 1, y_{10} receives 94% weight as compared to 6% weight for y_{13}; while the weights for r = 2 do not strongly indicate that both of these observations are outliers. Finally, in considering the analysis of variance in Table 9b, it is seen that the sums of squares corresponding to both L_1 and L_4 are large. This is, of course, what would be expected if it were known that any one of observations 10-13 was an outlier, since the effect that the bad observation would have on the posterior distribution of θ_{14} when $\alpha \neq 0$ is allowed would be "hidden" in the θ_{114} and θ_{144} third order bias coefficients under the restrictive $\alpha = 0$ model.

Computational Footnote.

The present speed and capacity of computers make it feasible to utilize model robustification techniques, such as those illustrated in these two examples, as well as in the example of section 4, involving John's 2^5 data, to guard against feared model discrepancies. However, much work needs to be done in the area of developing general programs which will actually do these robust analyses and thus take advantage of the existing computer capabilities.

References.

Andrews, D. F. (1971). A note on the selection of data transformations. Biometrika, 58, 249-54.

Bacon, D. W. (1970). Making the most of a one-shot experiment. Industrial and Engineering Chem., 62 (7), 27-34.

Box, G. E. P. (1979a). Robustness in the strategy of scientific model building. In Robustness in Statistics, 201-36, eds. Launer, R. L. and Wilkinson, G. N., New York, Academic Press.

Bailey, S. P. and Box, G. E. P. (1980). Modeling the nature and frequency of outliers. Tech. Summary Report # , Math. Research Center, Univ. of Wisconsin, Madison, Wisconsin.

Box, G. E. P. (1979b). Sampling and Bayes' inference in the advancement of learning. Tech. Report 1969, Math. Research Center, Univ. of Wisconsin, Madison, Wisconsin.

Box, G. E. P. (1979c). Unpublished expanded version of the previous reference.

Box, G. E. P. (1979d). Some problems of statistics and everyday life. J. Amer. Statist. Assoc., 74, 1-4.

Box, G. E. P. and Behnken, D. W. (1960b). Some new three-level designs for the study of quantitative variables. Technometrics, 2, 455-75.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). J. R. Statist. Soc. B, 26, 211-52.

Box, G. E. P. and Draper, N. R. (1979). Degrees of freedom, lack of fit, and transformation of predictors in composite designs. Unpublished manuscript.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. New York, Wiley.

Box, G. E. P. and Tiao, G. C. (1968). A Bayesian approach to some outlier problems. Biometrika, 55, 119-29.

Box, G. E. P. and Tiao, G. C. (1973). Bayesian Inference in Statistical Analysis. Reading, Mass., Addison-Wesley.

Box, G. E. P. and Tidwell, P. W. (1962). Transformation of the independent variables. Technometrics, 4, 531-50.

Chen. G. G. and Box, G. E. P. (1979a). Further study of robustification via a Bayesian approach. Tech. Report 1998, Math. Research Center, Univ. of Wisconsin, Madison, Wis.

Chen. G. G. and Box, G. E. P. (1979b). A study of real data. Tech. Report 2002, Math. Research Center, Univ. of Wisconsin, Madison, Wis.

Draper, N. R. and Herzberg, A. M. (1971). On lack of fit. Technometrics, 13, 231-41.

Draper, N. R. and Hunter, W. G. (1969). Transformations: Some examples revisited. Technometrics, 11, 23-40.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression I. Biometrika, 37, 409-28.

Durbin, J. and Watson, G. S. (1951). Testing for serial correlation in least squares regression II. Biometrika, 38, 159-78.

John, J. A. (1978). Outliers in factorial experiments. Appl. Statist. 27, 111-19.

John, J. A. and Draper, N. R. (1978). On testing for two outliers or one outlier in two-way tables. Technometrics, 20, 69-78.

Pallesen, L. C. (1978). Studies in the analysis of serially dependent data. Tech. Report 1837, Math. Research Center, Univ. of Wisconsin, Madison, Wis.

Stigler, S. M. (1977). Do robust estimators work well with real data? (with discussion.) Ann. Statist., 5, 1055-98.

Tukey, J. W. (1949). One degree of freedom for non-additivity. Biometrics, 5, 232-42.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2086	2. GOVT ACCESSION NO. AD-A089 643	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) THE DUALITY OF DIAGNOSTIC CHECKING AND ROBUSTIFICATION IN MODEL BUILDING: SOME CONSIDERATIONS AND EXAMPLES		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Steven P. Bailey and George E. P. Box		8. CONTRACT OR GRANT NUMBER(s) DAAG29-75-C-0024 DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit No. 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE June 1980
		13. NUMBER OF PAGES 49
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Diagnostic checks, robustification, discrepancy parameters, outliers, transformations, serial correlation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Consideration is given to the means by which appropriate diagnostic checking functions of the data can be developed to guard against feared model discrepancies. A formal basis for the selection of a function is given for situations where the feared inadequacy can be characterized by a discrepancy parameter β which takes a (possible inappropriate) value of β_0 in the model. The relationship of this checking function with the posterior distribution obtained from an elaborated ("robustified") model which allows for the discrepancy parameter to be estimated is discussed. The nature of the diagnostic check is briefly described

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract (continued)

for problems relating to the transformation of the dependent variable and to serial correlation; while a more thorough investigation of the checking function is given for problems relating to outlying observations and to transformation of the predictor variables. Several examples are given to illustrate these ideas.